

Accelerated Kaczmarz Algorithms using History Information

Tengfei Ma
IBM Research-Tokyo
feitengma0123@gmail.com

August 2, 2016

Abstract

The Kaczmarz algorithm is a well known iterative method for solving overdetermined linear systems. Its randomized version yields provably exponential convergence in expectation. In this paper, we propose two new methods to speed up the randomized Kaczmarz algorithm by utilizing the past estimates in the iterations. The first one utilize the past estimates to get a preconditioner. The second one combines the stochastic average gradient (SAG) method with the randomized Kaczmarz algorithm. It takes advantage of past gradients to improve the convergence speed. Numerical experiments indicate that the new algorithms can dramatically outperform the standard randomized Kaczmarz algorithm.

1 Introduction

The Kaczmarz algorithm ([Kaczmarz(1937)]) is a simple but powerful iterative method for solving the overdetermined system with equations $Ax = b$. Due to its simplicity and speed, it has a wide range of applications from computer tomography to image reconstruction ([Sezan and Stark(1987)]). It is a form of alternating projection method, which in each iteration projects the current solution to a subspace.

Given a matrix $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ and $b \in \mathbb{R}^m$, we denote the rows of A by $a_1^T, a_2^T, \dots, a_m^T$ and $b = (b_1, b_2, \dots, b_m)^T$. The Kaczmarz method project the current estimation orthogonally onto the solution hyperplane of $a_j^T x = b_j$, where the row j is selected in a cyclic manner.

Recently, [Strohmer and Vershynin(2009)] proposed to select the row with biased sampling and proved that the randomized Kaczmarz method (RK) converges with expected exponential rate. Let $\|A\|_F^2$ denote the Frobenius norm of A and $\|\cdot\|$ denote the standard norm. In each iteration a row i is randomly selected with probability proportional to $\|a_i\|^2$, and finally we could get the following exponential bound for the convergence in expectation:

$$\mathbb{E}\|x_k - x\|^2 \leq \left(1 - \frac{1}{\kappa(A)^2}\right)^k \|x_0 - x\|^2 \quad (1)$$

where $\kappa(A) = \|A\|_F \|A^{-1}\|$, A^{-1} is the left inverse of A which is always assumed to exist, and x_0 is an arbitrary initial value. Using this algorithm, the cost per iteration is $O(n)$ and the expected iteration for convergence is $O(\log(1/\epsilon))$ where ϵ is the accuracy parameter. So it is computationally feasible for very large systems. It is also shown that the randomized Kaczmarz method often outperforms the celebrated conjugate gradient method.

Besides consistent linear system, the RK algorithm has also been analyzed for inconsistent linear system $Ax = b + w$ where w is an arbitrary noise vector ([Needell(2010)]). And some extended RK methods are proposed for systems of linear inequalities ([Leventhal and Lewis(2010)]), least square problems ([Zouzias and Freris(2013)]), and online compressed sensing ([Lorenz et al.(2014)Lorenz, Wenger, Schopfer, Magnor, et al.]). The RK algorithm has a theoretical linear convergence rate. However, the convergence rate largely depends on the condition number κ of matrix A , and the convergence will be extremely slow for ill-conditioned problems. Therefore, some accelerated RK methods are proposed. For example, [Liu and Wright(2015)] applied the Nesterov acceleration scheme to the standard RK algorithm, and obtained the accelerated randomized Kaczmarz algorithm (ARK).

In this paper, we develop new acceleration schemes for the Kaczmarz algorithm by utilizing history information. The basic idea is to change the direction of projection in each iteration to make it converge faster. In the RK algorithm, each projection is always along a row vector a_j which is orthogonal to the hyperplane $a_j^T x = b_j$. Our first acceleration scheme finds a new preconditioner C which changes the projection direction into Ca_j . The preconditioner is approximated based on the estimate of x in past iterations. Our second acceleration scheme considers the relationship between stochastic gradient descent (SGD) and the randomized Kaczmarz algorithm and combines them. In each iteration, we first use the past gradients to get a variant of SGD, the stochastic average gradient (SAG). We do a gradient descent along the SAG and then project the point back into a hyperplane.

The paper is organized as follows. The next section covers related work about accelerated Kaczmarz algorithms. In section 3, we introduce the preconditioning technique and induce our new preconditioner. Section 4 present the second acceleration scheme which integrates SAG into the RK algorithm. Numeric experiments are shown in Section 5 and we conclude the paper in Section 6.

Algorithm 1 Randomized Kaczmarz Algorithm

- 1: Initialize $k \leftarrow 0$
 - 2: **for** $k = 0, 1, \dots$ **do**
 - 3: Select row j from $\{1, 2, \dots, m\}$ with probability $\frac{\|a_j\|^2}{\|A\|_F^2}$
 - 4: Project $x_{k+1} = x_k + \frac{(b_j - a_j^T x_k)}{\|a_j\|^2} a_j$
 - 5: Update $k \leftarrow k + 1$
 - 6: **end for**
-

2 Related Work

2.1 Improvement to the Kaczmarz algorithm

Since the RK was analyzed by [Strohmer and Vershynin(2009)], there has been several directions to extend it. Two-subspace projection method extends RK by iterately projecting the estimate onto the solution space given by two randomly selected rows. It only improves when the system has correlated rows. [Eldar and Needell(2011)] has a different strategy to select the rows. It projects the row vectors onto a low dimensional space, and then selects the row which leads to the largest improvement. Beyond the scope of randomized Kaczmarz algorithm, there is some work focusing on accelerating the classical Kaczmarz method. For example, Brezinski and RedivoZaglia [Brezinski and Redivo-Zaglia(2013)] use sequence transformation to change the projection procedure. But it is too complex to get a transformed sequence and it needs to store too many vectors, so it is difficult to be applied in practice when the dimension of A is large.

2.2 The Randomized Kaczmarz Algorithm and Stochastic Gradient Descent

The Kaczmarz algorithm is fundamentally a special case of alternating projection ([Strohmer and Vershynin(2009)]). In some area it is called POCS (projection to convex sets). But it also has a strong relationship with stochastic gradient descent (SGD). Very recently, It has been demonstrated that the RK algorithm is equivalent to a form of stochastic gradient descent with weighted sampling ([Needell et al.(2014)Needell, Ward, and Srebro]). However, one advantage of the RK is that it does not need to set up the step size for each iteration, although it generally does not get the optimal step size.

Considering the connection between the RK and the SGD, Liu and Stephen [Liu and Wright(2015)] apply the well known Nesterov's acceleration procedure to the RK algorithm. They demonstrate the convergence of their accelerated randomized kaczmarz algorithm (ARK) and obtain significant improvement for ill-conditioning problems in numeric experiments ([Liu and Wright(2015)]). The ARK introduces two additional sequences $\{y_k\}$ and $\{v_k\}$ as follows

$$\begin{aligned} y_k &= \alpha_k v_k + (1 - \alpha_k) x_k \\ x_{k+1} &= y_k - a_i(a_i^T y_k - b_i)/\|a_i\|^2 \\ v_{k+1} &= \beta_k v_k + (1 - \beta_k) y_k - \gamma a_i(a_i^T y_k - b_i)/\|a_i\|^2 \end{aligned}$$

where the scalars α_k, β_k and γ_k are calculated offline based on the hyperparameter $\lambda \in [0, \lambda_{min}]$, (λ_{min} is the minimum eigenvalue of $A^T A$). They prove that when $\lambda > 0$, the ARK gets a linear convergence rate. The ARK is then extended to solving the sparse data.

The ARK performs very well for ill-conditioned problems, especially when the $\lambda_{min}(A^T A)$ is known. However, to get the accurate $\lambda_{min}(A^T A)$ is difficult. And in many cases the inaccurate $\lambda_{min}(A^T A)$ will lead to much worse performance.

2.3 Importance of History Information

History information has been used in many acceleration schemes for alternating projection ([Gearhart and Koshy(1989)]) and SGD ([Roux et al.(2012)Roux, Schmidt, and Bach], [Johnson and Zhang(2013)], [Nitanda(2014)]). The main idea to use history information is to find a point closest to the final solution in each iteration ([Gearhart and Koshy(1989)]) or reduce the variance between stochastic gradients and the full gradients ([Johnson and Zhang(2013)]). The ARK also utilizes the history information by employing a Nesterov's acceleration procedure. In this paper, we developed two approaches to utilizing history information. We use the past estimates of x to approximate a preconditioner in our first algorithm, while in the second algorithm we use the past stochastic gradients to approximate full gradients as in ([Roux et al.(2012)Roux, Schmidt, and Bach]).

3 Approximated Preconditioned Kaczmarz(APK) Algorithm

The motivation of our first acceleration scheme lies on two aspects. Firstly, we consider using preconditioning to reduce the condition number of A in the system. Secondly, we want to use the history information to generate a proper preconditioner.

As we explained before, the convergence rate of the randomized Kaczmarz algorithm largely depends on the condition number of A . When this number is large, the convergence speed will be too slow. One solution to this problem is to use a preconditioning matrix. A preconditioning matrix (or preconditioner) B of a matrix A is a matrix such that BA or AB has a smaller condition number than A . So the original problem could be changed into either a left preconditioned system

$$BAx = Bb$$

or a right preconditioned system.

$$ABB^{-1}x = b \tag{2}$$

Here we consider the right preconditioned system 2. Assume that we already know the preconditioner $B \in \mathbb{R}^{n \times n}$, we explain each iteration of the new Kaczmarz algorithm as follows.

First we solve the new linear system $AB y = b$, where $y = B^{-1}x$. So at each iteration we project the current estimation y_k on the hyperplane defined by the row i :

$$y_{k+1} = y_k + \frac{b_i - a_i^T B y_k}{a_i^T B B^T a_i} B^T a_i \tag{3}$$

Replace y_k with $B^{-1}x_k$, then we get the update of x :

$$B^{-1}x_{k+1} = B^{-1}x_k + \frac{b_i - a_i^T B B^{-1}x_k}{a_i^T B B^T a_i} B^T a_i \tag{4}$$

$$x_{k+1} = x_k + \frac{b_i - a_i^T x_k}{a_i^T B B^T a_i} B B^T a_i \tag{5}$$

A good preconditioner may accelerate the convergence a lot. Indeed, the choice of preconditioner is often more important than the choice of iterative method, according to Yousef Saad ([Saad(2003)]). However, how to select a preconditioning matrix remains a difficult problem. In many cases determining a good preconditioning matrix itself has the same computational complexity with the original problem.

In order to use a preconditioner in the Kaczmarz algorithm, it is better to keep the preconditioner to be a diagonal matrix. Each iteration of the Kaczmarz algorithm costs only $O(n)$. If the preconditioner B is not diagonal or sparse enough, the computation of $a_i^T B$ costs $O(n^2)$, which will be unacceptable for just one iteration. Another choice is to directly get $BB^T a_i$ without an previous estimation of B , such as the online LBFS (oLBFS) method ([Schraudolph et al.(2007)Schraudolph, Yu, and Günter]), which is a stochastic quasi-newton method. But it still costs more than using a diagonal matrix.

We aim to get a diagonal matrix $C = BB^T$, thus the "projection" could be represented by a modified form which only contains C :

$$x_{k+1} = x_k + \frac{b_i - a_i^T x_k}{a_i^T C a_i} C a_i \quad (6)$$

Left-multiply the two sides with a vector a_i^T , we find that $a_i^T x_{k+1} = b_i$. That means x_{k+1} is still on the hyperplane given by the original row i .

3.1 Optimizing the diagonal preconditioner using history information

Consider that the Kaczmarz algorithm is essentially an alternating projection method. After each iteration, the new estimation lay on a hyperplane.

Assume that we have a list of past estimates in the classical Kaczmarz algorithm $x_1, \dots, x_m, x_{m+1}, x_{2m}$, where rows are selected with a cyclic manner according to an order $R(1, \dots, m)$. So each x_k is projected onto the hyperplane given by the row $i_{k+1} = R(k+1)$ and leads to the next estimation x_{k+1} . Since we do not change the selection order, x_k and x_{k+m} are on the same hyperplane.

The idea of our method is, why do not we directly use a pseudo projection to project x_k to x_{k+m+1} instead of the original x_{k+1} ? That means, we let the pseudo projection seemingly jump across a cycle of projections. So we use the preconditioned row vector $a_{i_k} C$ as the direction of the pseudo projection. we show a very simple case in Figure 1 as an example.

$$C = \arg \min_C F(C) = \arg \min_C \sum_{k=2}^m \|x_{k+m} - x'_k\|^2 \quad (7)$$

where x'_k is a projection of x_{k-1} along the direction of $a_{i_k} C$ instead of a_{i_k} . So the objective function becomes

$$F_1(C) = \sum_{k=2}^m \left\| x_{k+m} - x_{k-1} - \frac{b_{i_k} - a_{i_k}^T x_{k-1}}{a_{i_k}^T C a_{i_k}} C a_{i_k} \right\|^2 \quad (8)$$

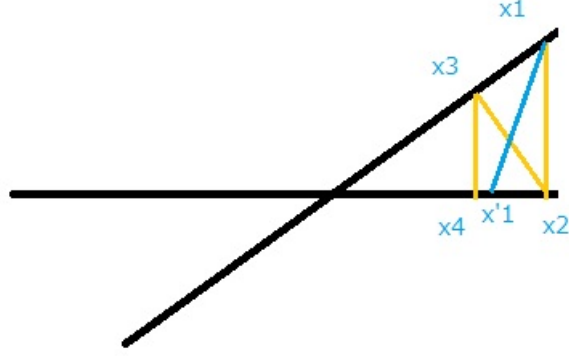


Figure 1: A simple example of the preconditioning idea. In this simple $\mathbb{R}^{2 \times 2}$ case, we want to use a new projection vector. It projects x_1 to x'_1 which is close to x_4

To simplify the optimization problem, we have the following strategy to approximate the objective function. We assume that C is not too distant from I , so that we could have $a_i^T C a_i \simeq a_i^T a_{i_k}$. In this case, the objection function has been changed into a combination of two parts, an approximation of the Equation8, and a regularization term to keep $a_{i_k}^T C a_{i_k}$ close to $a_{i_k}^T a_{i_k}$. To keep similarity, we used the Frobenius norm of (C-I) as regularization.

$$\sum_{k=2}^m \left\| x_{k+m} - x_{k-1} - \frac{b_{i_k} - a_{i_k}^T x_{k-1}}{a_{i_k}^T a_{i_k}} C a_{i_k} \right\|^2 + \alpha \|C - I\|_F^2 \quad (9)$$

where α is a coefficient parameter for the regularization.

As C is diagonal, we only need to calculate the diagonal vector of C . Denote $s = \text{diag}(C)$ as the diagonal vector of C , and A_{i_k} as a diagonal matrix whose diagonal is a_{i_k} . Then $C a_{i_k}$ could be transformed into $A_{i_k} s$, and $F_2(C)$ can be written as functions of s : $F_2(C) = F(s)$.

As the objective function turned to be convex, it is easy to get the solution by making the derivative $F'(s) = 0$. From the objective function

$$F(s) = \sum_{k=2}^m \left\| x_{k+m} - x_{k-1} - \frac{b_{i_k} - a_{i_k}^T x_{k-1}}{a_{i_k}^T a_{i_k}} A_{i_k} s \right\|^2 + \alpha \|(s - e)\|^2 \quad (10)$$

where e is a all-ones vector, we differentiate $F(s)$ with respect to s , and get the derivative as

$$F'(s) = F'_1(s) + 2\alpha(s - e) \quad (11)$$

where

$$F_1'(s) = -2 \sum_{k=2}^m \left(\frac{b_{i_k} - a_{i_k}^T x_{k-1}}{a_{i_k}^T a_{i_k}} \right) A_{i_k} (x_{k+m} - x_{i_k} - \frac{b_{i_k} - a_{i_k}^T x_{k-1}}{a_{i_k}^T a_{i_k}} A_{i_k} s) \quad (12)$$

Let $F_1'(s) = 0$, $A'_{i_k} = \frac{b_{i_k} - a_{i_k}^T x_{k-1}}{a_{i_k}^T a_{i_k}} A_{i_k}$, and $\delta_{i_k} = x_{k+m} - x_{k-1}$, we get the optimal s :

$$\begin{aligned} s &= \arg \min_s F_1(s) \\ &= \left(\sum_{k=2}^m (\alpha + A_{i_k}'^2) \right)^{-1} \left(\sum_{k=2}^m A_{i_k}' \delta_{i_k} + \alpha e \right) \end{aligned} \quad (13)$$

Note that A'_{i_k} is a diagonal matrix, so the computation costs only $O(mn)$. And when m is extremely large, we can use only a subset of samples from 1,... m instead of a full computation. In practice we update this diagonal matrix after a long interval, so the costs does not impact a lot.

3.2 Convergence analysis of the APK Algorithm

Proposition. 1 *The APK algorithm converges at a linear rate.*

The conclusion is very intuitive. It can be easily proved from the convergence analysis of the randomized Kaczmarz algorithm 1. We apply the conclusion (1) to the right preconditioned form (2) which we use in our APK algorithm:

$$\mathbb{E} \|B^{-1}x_k - B^{-1}x\|^2 \leq \left(1 - \frac{1}{\kappa(AB)^2}\right)^{k-t_0} \|x_{t_0} - x\|^2 \quad (14)$$

where x_{t_0} is the start point after we update the preconditioner C , and x_k is an estimate before we update C next time. Denote the minimum and maximum eigenvalues of $C = B * B$ as $\lambda_{min}(C)$ and $\lambda_{max}(C)$ separately. Then $\|B^{-1}x_k - B^{-1}x\|^2 \geq \lambda_{max}(C)^{-1} \|x_k - x\|^2$. So we could get the convergence rate of the ARK algorithm:

$$\mathbb{E} \|x_k - x\|^2 \leq \lambda_{max}(C) \left(1 - \frac{1}{\kappa(AB)^2}\right)^{k-t_0} \|x_{t_0} - x\|^2 \quad (15)$$

3.3 Other Preconditioners

Solving a linear system $Ax = b$ is equivalent to solving the least square problem $\sum_i \|b_i - a_i^T x\|^2$. When we use a stochastic gradient descent method for this problem, a good preconditioner is the inverse Hessian matrix. And for computational efficiency, some methods have been proposed to approximate the Hessian matrix by a diagonal one, i.e. diagonal Hessian matrix. One state-of-art method is the AdaGrad method ([Duchi et al.(2011)Duchi, Hazan, and Singer]).

3.3.1 AdaGrad

The motivation of the AdaGrad is to incorporate the geometry knowledge of the data observed in earlier iterations to adapt the weights of each dimension. At each step t , we receive a subgradient $g_t \in \partial f_t(x_t)$ of f_t at x_t . Update $g_{1:t} = [g_{1:t-1} \ g_t]$, $s_{t,i} = \|g_{1:t,i}\|$.

Then the diagonal Hessian matrix is approximated by

$$H_t = \zeta I + \text{diag}(s_t). \quad (16)$$

We can use the inverse of Hessian as our preconditioner. In practice, we find that it is better to add another decay term for the inverse Hessian approximation. So, finally we use $C = \lambda_0 + H_t^{-1}$. In section 6, we compare the results of our APK algorithm and AdaSGD.

4 The Stochastic Average Gradient based Randomized Kaczmarz Algorithm

As we introduced before, the randomized Kaczmarz algorithm can be regarded as a special form of stochastic gradient descent ([Needell et al.(2014)Needell, Ward, and Srebro]). Solving a linear system $Ax = b$ is equivalent to solving the least square problem $\sum_i f_i = \sum_i \|b_i - a_i^T x\|^2$. In the randomized Kaczmarz algorithm 1, the $(b_j - a_j^T x_k)a_j$ is the gradient of a component f_j , and $\frac{1}{\|a_j\|^2}$ can be seen as the step size. The connection between RK and SGD motivates us to bring in acceleration schemes of the SGD into the RK algorithm.

4.1 Stochastic Average Gradient (SAG)

Recently, there has been a lot of work on accelerating the SGD, such as SAG ([Roux et al.(2012)Roux, Schmidt, and Bach]), SDCA ([Shalev-Shwartz and Zhang(2013)]), SVRG ([Johnson and Zhang(2013)]), SAGA ([Defazio et al.(2014)Defazio, Bach, and Lacoste-Julien]). The SAG algorithm is one of the simplest of them. It requests only a small number of operations at each iteration.

As in a general stochastic optimization, the SAG method typically solve the problem of optimizing a sum of functions in this form:

$$\min_{x \in \mathbb{R}_n} g(x) := \frac{1}{m} \sum_{i=1}^m f_i(x)$$

where each f_i is convex and each gradient f'_i is Lipschitz continuous with constant L . To get a linear convergence rate, the average function $g(x)$ is also assumed strongly convex.

- f'_i is Lipschitz continuous:

$$\|f'_i(x) - f'_i(y)\| \leq L\|x - y\|$$

- g is strongly convex:

$$g(x) \geq g(y) + g'(y)(x - y) + \frac{\mu}{2} \|x - y\|^2$$

The SAG method combines the low iteration cost of the stochastic gradient descent methods with a linear convergence rate as in the full gradient methods. The method stores the most recent gradient of f_i ($i = 1, \dots, m$) and use the average of them to approximate the full gradient vector. At each iteration, a random training example i_k is selected and x is updated:

$$x^{k+1} = x^k - \frac{\alpha_k}{m} \sum_{i=1}^m \phi_i^k$$

where ϕ_i^k is updated as follows:

$$\phi_i^k = \begin{cases} f'_i(x^k) & \text{if } i = i_k \\ \phi_i^{k-1} & \text{otherwise} \end{cases}$$

Let $d = \frac{1}{m} \sum_{i=1}^m \phi_i^k$, then at each procedure we only need to update d by the following procedure:

$$d_{k+1} = d_k - \phi_{i_k} + f'_{i_k}(x^k)$$

The SAG method essentially reduced the variance between the stochastic average gradient and the full gradient ([Defazio et al.(2014)Defazio, Bach, and Lacoste-Julien]). A variance reduction approach is to use $\alpha(X - Y) + \mathbb{E}Y$ as an approximation of $\mathbb{E}X$, where $\alpha \in [0, 1]$, X is the SGD gradient, and Y is the past stored gradient. So the variance could be changed from $Cov(X, Y)$ to $\alpha[Var(X) + Var[Y] - 2Cov(X, Y)]$. SAG could be obtained from the technique by using $\alpha = 1/n$. Using the same form, when $\alpha = 1$, we could get the SAGA, which is unbiased but has larger variance than SAG. Most recently, non-uniform version of SAG, namely SAG-NUS, has also been developed to generalize SAG and SAGA, where $\alpha = \frac{1}{np}$.

SAG, SAGA, and SAG-NUS all use a constant step size in each iteration. This makes the algorithms converge fast while they are also easy to be implemented. For example, it can be demonstrated that with a constant step size of $\alpha_k = \frac{1}{2mL}$, the SAG iterations satisfy

$$\mathbb{E}[\|x^k - x^*\|^2] \leq (1 - \frac{\mu}{8Lm})^k [3\|x_0 - x^*\|^2 + C_0] \quad (17)$$

where $C_0 = \frac{9\sigma^2}{4L^2}$, and σ is the variance of the gradient norms at the final solution x^* . Another choice $\alpha_k = \frac{1}{16L}$ has also been proved convergent in a later version of the SAG work ([Schmidt et al.(2013)Schmidt, Roux, and Bach]). The SAGA method and the SAG-NUS select larger step sizes which could theoretically guarantee linear convergence. However, [Schmidt et al.(2013)Schmidt, Roux, and Bach] indicate that in practice we could also select a larger step size for SAG. They gave two recommendations for the step size: $1/L$ and $2/(L + m\mu)$, and observed that $1/L$ always converged and performs better than the step size of $1/16L$.

4.2 SAG-RK

SAG is very efficient when the objective function is strongly convex. However, in a linear system, each component of the objective function $\|b_i - a_i^T x\|^2$ is not strongly convex if we do not add extra regularizations. In contrast, the random Kaczmarz does not have this limitation. Our second acceleration scheme combines the methods of SAG and RK. First, we use the stochastic gradient descent to make a descent direction. Then we project the point back onto the hyperplanes of each row in the linear system.

Algorithm 2 The SAG-RK Algorithm

1: **for** $k = 0, 1, \dots$ **do**

2: Select a row j from $\{1, 2, \dots, m\}$ with probability $\frac{\|a_j\|^2}{\|A\|_F^2}$

3: Calculate the stochastic average descent g_k

4: SAG descent step:

$$y_k = x_k - \alpha_k g_k$$

5: Project

$$x_{k+1} = y_k + \frac{(b_j - a_j^T y_k)}{\|a_j\|^2} a_j$$

6: Update history information

7: Update $k \leftarrow k + 1$

8: **end for**

The idea is motivated by the incremental constraint projection-proximal methods [?], which extends the projection/proximal gradient methods by using random subgradient and random constraint updates. Given a convex optimization problem $\min_{x \in X} \sum_i^N f_i(x)$ with constraints $X = \cap_{j=1}^m X_j$, the algorithm updates $x_{k+1} = \Pi_j[x_k - \alpha_k g_i(x_k)]$ in each iteration by sampling a j from the constraints and an i from the components of $\sum_i^N f_i(x)$, where $g_i(x_k)$ indicates the subgradient of $f_i(x)$ at x_k , Π_j denotes a Euclidean projection onto X_j .

In our problem we rewrite our problem as $\min_{x \in X} \sum_i \|b_i - a_i^T x\|^2$ and the constraints are $X = \cap_{j=1}^m (a_j^T x = b_j)$. For the minimum objection, we use SAG in each iteration, and then project the point onto a hyperplane $a_j^T x = b_j$.

SAG-RK also has a strong relationship with SAG-NUS. From the update equation, we get

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k g_k + \frac{(b_j - a_j^T y_k)}{\|a_j\|^2} a_j \\ &= x_k - \alpha_k g'_k \end{aligned}$$

where $g'_k = g_{k-1} + \beta(-\phi_i + f'_i(x^k))$, $\beta = 1/m - 1/\alpha_k \frac{(b_j - a_j^T y_k)}{a_j^T (y_k - y')}$, and y' is the last estimate associated with ϕ_j . So the SAG-RK is in the same family as SAG-NUS, SAG and SAGA. They only differentiate by using a different weight β .

In practice, if the step size α_k is small enough, we could use $x_{k+1} = y_k + \frac{(b_j - a_j^T x_k)}{\|a_j\|^2} a_j$ instead to update x in the algorithm. This change could eliminate the calculation of gradient at y_k and decrease the computational time. In fact, it is equivalent to adding a relaxation parameter to the projection step, so we call this implementation SAG-RK-relaxation. In the next section, we will compare the two implementations.

5 Numerical Experiments

In this section, we study the computational behavior of APK and SAG-RK, and compare them with the original RK and other acceleration schemes: the AdaGrad and the ARK.

5.1 Synthetic Data

We adopt two strategies of generating synthetic data which are used in RK ([Strohmer and Vershynin(2009)]) and ARK ([Liu and Wright(2015)]) separately. For overdetermined system ($m > n$), we follow [Strohmer and Vershynin(2009)] and let A be a $m * n$ matrix whose entries are independent $N(0, 1)$ random variables. In this case, the condition number of A converges to:

$$\frac{\kappa(A)}{\sqrt{n}} \rightarrow \frac{1}{1 - \sqrt{m/n}}$$

As m/n decrease, the condition number becomes larger and larger. So, when $m = n$, we cannot control the large condition number by this method. Then we use the method in [Liu and Wright(2015)] instead. We first generate a random $n * n$ Gaussian matrix and find its SVD: $U\Lambda V^T$. Next, we change the singular values to $\Lambda_{ii} = i^{-\alpha}$ and compute $U\Lambda V^T$ again to get a new A .

For the overdetermined case, we use $m = 500, n = 400$; and for a square matrix, we use $m = n = 500$ and generate two matrices with $\alpha = 0.75, 0.9$ separately.

5.2 Implementation

- **AdaGrad:** Using Equation (16) to get an approximate diagonal Hessian matrix, then we add a decay term to its inverse matrix as the preconditioner $C = \lambda_0 + H_t^{-1}$ for our linear system (In the experiments, we set $\lambda_0 = 0.2$ which gets the best performance). This matrix C is then employed for 6 and get a AdaGrad based RK algorithm.
- **ARK:** As we mentioned before, to implement the ARK algorithm, we have to estimate the parameter λ_{min} first. In Liu and Wrigt's work ([Liu and Wright(2015)]), they have a strategy to approximate the real λ_{min} . Run RK for $K2$ iterations and record x_{K2+1} and x_{K1+1} where $K1 = \max(1, K2 - 10m)$. Based on the con-

vergence rate, they estimate the λ_{min} as follows¹:

$$\|A\|_F^2 \left[1 - \left(\frac{\|Ax_{K2} - b\|}{\|Ax_{K1} - b\|} \right)^{\frac{0.5}{K_2 - K_1}} \right]$$

So the ARK needs a long burn-in time to determine a good parameter. In this paper we fix $K_2 = 15m$.

For our own two algorithms, we follow the procedures that are introduced in Section 3 and Section 4. Selecting a proper step size is important for SAG, so we try different choices and compare them first. In Figure 2(a), we show residual errors for SAG-RK and its relaxation implementation with two stepsizes separately on a 500×400 matrix. At each stepsize, the two implementations have almost the same performance, and a larger step size leads to significantly faster convergence².

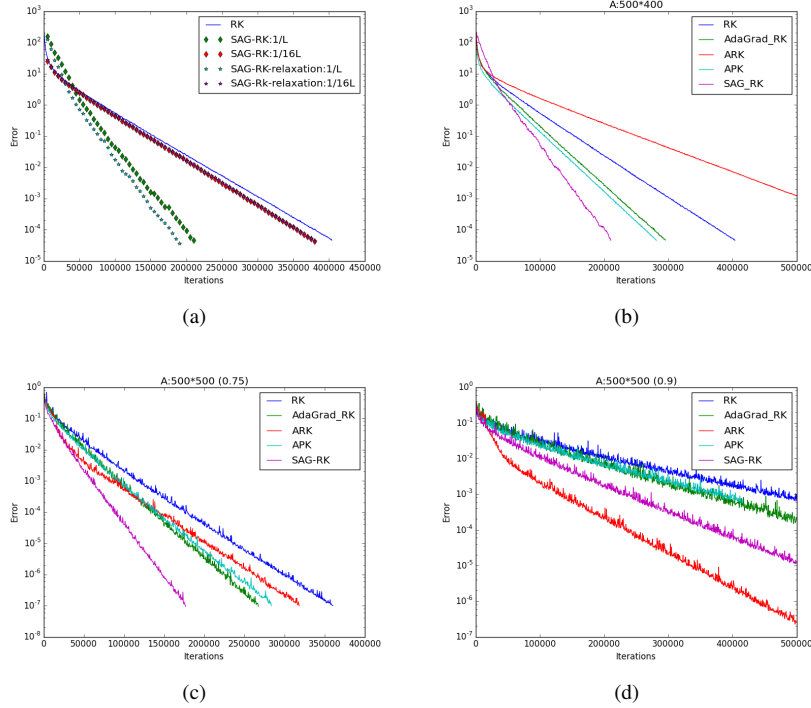


Figure 2: Numeric Experiments. (a): Comparison of step sizes in SAG-RK. (b): $A_1 \in \mathbb{R}^{500 \times 400}$ and $\kappa(A_1) = 180.7$ (c): $A_2 \in \mathbb{R}^{500 \times 500}$, and $\kappa(A_2) = 167.9$ (d): $A_3 \in \mathbb{R}^{500 \times 500}$, and A has a larger condition number $\kappa(A_3) = 367.6$

¹In the original paper, they normalize the matrix A so that $\|A\|_F^2 = m$

²It has a different story when $m \gg n$, but it is beyond the scope of this paper, because we only consider the ill-conditioned case.

Table 1: Computational time (seconds) until $\|b - Ax\|/\|b\| < 10^{-7}$

| Model | A_1 | A_2 | A_3 |
|--------------|--------|-------|--------|
| RK | 50.80 | 43.42 | 159.17 |
| SAG | 339.29 | 48.25 | 182.65 |
| SAG-RK | 36.87 | 31.35 | 116.13 |
| SAG-RK2 | 31.95 | 26.87 | 99.63 |
| APK | 38.41 | 40.43 | 156.86 |

Then we choose the step size of SAG-RK as $1/L$, and show the performance of all algorithms in Figure 2(b), 2(c), 2(d). First, we compare the APK and AdaGrad, the APK has only slightly better convergence rates over AdaGrad. However, AdaGrad needs to update the diagonal matrix at each iteration and costs too much for the RK algorithm, while APK only update the preconditioner after a long interval. So APK is much more computationally efficient.

Then we compare our algorithm with the ARK algorithm. The ARK algorithm does not perform well when the condition number is not large enough, it is even worse than the RK algorithm. But it has the best performance when we have the largest condition number. This agrees with the conclusions in [Liu and Wright(2015)] that the ARK only suits to seriously ill-conditioned problems. In contrast to ARK, our algorithms performs consistently better than RK. In particular, the SAG-RK performs best on A_1 and A_2 and second best on A_3 .

At last, we show how our algorithms improve RK on computational time in Table 1. As SAG-RK can be seen a combination of SAG and RK, we compare it with the two algorithms. As we discussed before, to save the operation time at each iteration, we use another implementation, namely SAG-RK-relaxation (In the table, we call it SAG-RK2 to save space). All algorithms check the residual error every $10m$ iterations and stop at $\|b - Ax\|/\|b\| < 10^{-7}$.

From Table 1, we could see that SAG-RK has a much shorter computational time than RK and SAG. Furthermore, the SAG-RK-relaxation indeed improves the computational efficiency and outperforms all others. APK also has a descent performance, but it cannot get significant improvement over RK in all cases.

6 Conclusion

In this paper, we propose two methods to accelerate the randomized Kaczmarz algorithm, namely APK and SAG-RK. They both take advantage of the history information in past iterations. APK use past estimates of x to get an approximate right preconditioner for the linear system, while SAG-RK use past gradients to get an approximate full gradient and combine the SAG step with a Kaczmarz projection. The APK provides a new vision to develop preconditioners based on history information. And the performance of SAG-RK exceeds both SAG and RK consistently. Future work includes

extension to the inconsistent linear systems as well as the sparse case, and theoretic demonstration of the convergence rate.

References

- [Brezinski and Redivo-Zaglia(2013)] C. Brezinski and M. Redivo-Zaglia. Convergence acceleration of kaczmarz method. *Journal of Engineering Mathematics*, pages 1–17, 2013.
- [Defazio et al.(2014)Defazio, Bach, and Lacoste-Julien] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.
- [Duchi et al.(2011)Duchi, Hazan, and Singer] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.
- [Eldar and Needell(2011)] Y. C. Eldar and D. Needell. Acceleration of randomized kaczmarz method via the johnson–lindenstrauss lemma. *Numerical Algorithms*, 58(2):163–177, 2011.
- [Gearhart and Koshy(1989)] W. B. Gearhart and M. Koshy. Acceleration schemes for the method of alternating projections. *Journal of Computational and Applied Mathematics*, 26(3):235–249, 1989.
- [Johnson and Zhang(2013)] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.
- [Kaczmarz(1937)] S. Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bulletin International de l’Academie Polonaise des Sciences et des Lettres*, 35:355–357, 1937.
- [Leventhal and Lewis(2010)] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010.
- [Liu and Wright(2015)] J. Liu and S. Wright. An accelerated randomized kaczmarz algorithm. *Mathematics of Computation*, 2015.
- [Lorenz et al.(2014)Lorenz, Wenger, Schopfer, Magnor, et al.] D. Lorenz, S. Wenger, F. Schopfer, M. Magnor, et al. A sparse kaczmarz solver and a linearized bregman method for online compressed sensing. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 1347–1351. IEEE, 2014.
- [Mairal(2013)] J. Mairal. Optimization with first-order surrogate functions. In *Proceedings of The 30th International Conference on Machine Learning*, pages 783–791, 2013.

- [Needell(2010)] D. Needell. Randomized kaczmarz solver for noisy linear systems. *BIT Numerical Mathematics*, 50(2):395–403, 2010.
- [Needell et al.(2014)Needell, Ward, and Srebro] D. Needell, R. Ward, and N. Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in Neural Information Processing Systems*, pages 1017–1025, 2014.
- [Nitanda(2014)] A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems*, pages 1574–1582, 2014.
- [Roux et al.(2012)Roux, Schmidt, and Bach] N. L. Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.
- [Saad(2003)] Y. Saad. *Iterative methods for sparse linear systems*. Siam, 2003.
- [Schmidt et al.(2013)Schmidt, Roux, and Bach] M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *arXiv preprint arXiv:1309.2388*, 2013.
- [Schraudolph et al.(2007)Schraudolph, Yu, and Günter] N. N. Schraudolph, J. Yu, and S. Günter. A stochastic quasi-newton method for online convex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 436–443, 2007.
- [Sezan and Stark(1987)] M. I. Sezan and H. Stark. Applications of convex projection theory to image recovery in tomography and related areas. *Image Recovery: Theory and Application*, pages 155–270, 1987.
- [Shalev-Shwartz and Zhang(2013)] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599, 2013.
- [Strohmer and Vershynin(2009)] T. Strohmer and R. Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.
- [Zouzias and Freris(2013)] A. Zouzias and N. M. Freris. Randomized extended kaczmarz for solving least squares. *SIAM Journal on Matrix Analysis and Applications*, 34(2):773–793, 2013.